HONAM MATHEMATICAL JOURNAL
Volume 4, Number 1, July. 1982

Estimating Missing Points In Experiments

By JUNG WOOK SIM

Dept. of Math. College of Natural Sciences, Chonnam National Univ.

實驗에 있어서 缺測点 推定

沈 政 煜

全南大學校 自然科學大學 數學科

Summary

Estimation methods of missing points for an experimental design are described. Formulae are provided for the estimation of missing points using matrix notation. The correct analysis of variance table is given. Estimation methods of a single missing point and two missing points in 2^n factorial designs are described.

1. Introduction

The need for estimating missing points has long been recognized by statisticians dealing with experimental design. For instance, a simple randomized block design containing one or more missing values requires a relatively complicated least squares analysis; the re-introduction of the missing points restores the balance of the design, and hence the easier analysis.

Allen and Wishart (1930) provided formulae appropriate in the case of a single plot of a randomized block or Latin square. Yates (1933) suggested an iterative procedure to deal with more than one missing point. Both of these methods for qualitative designs minimize the error term in the analysis of variance. They can also be used to estimate missing values in a full factorial design.

However a large proportion of industrial planned experiments use qualitative designs which are not fully replicated, and such designs have not been catered for with similar simple methods to the above. Wilkinson (1958, 1970) and Tocher (1952) provided examples of methods which can require considerable computation

owing to matrix manipulation and inversion.

A simpler technique for factorial designs which does not minimize the residual sum of squares, was suggested by Draper and Stoneman (1963). The methed is to equate a certain number of high-order interactions to zero, and check on lack of bias in the lower effects by means of a half normal plot. In the case of one missing point estimation proceeds by equating the highest order interaction to zero - this is equivalent to minimizing the residual sum of squares where the residual is estimated by the highest order interaction. More than one missing value is dealt with by zeroing an equal number of interactions.

2. Estimation of Missing points

Suppose we fit by least squares, to the results $y'=(y_1, y_2, \cdots, y_N)$ of an experimental investigation, a regression equation of the form $\hat{\underline{y}}=X\underline{b}$, *i*, *e*., the model considered is $\underline{y}=X\underline{\beta}+\underline{e}$, where $\underline{e}\sim N(O, I\sigma^2)$. Then $\underline{b}=(X'X)^{-1}X'\underline{y}$.

Suppose that the matrix X' is divided into $[X_1', X_2']$ in such a way that X_1 is associated with yield values $\underline{y}_1$ that are observed, and X_2 is associated with yield values $\underline{y}_2$ that are missing. This is easily effected by rearranging the order of the symbols $y_1, y_2, \cdots, y_N$ so that the h missing values occupy the last h places $y_{(N-h+1)}, \cdots, y_N$ and rearranging the rows of X to correspond so that the model remains the same,

$$E(\underline{y})=E\begin{pmatrix}\underline{y}_1\\ \underline{y}_2\end{pmatrix}=\begin{pmatrix}X_1\\ X_2\end{pmatrix}\underline{\beta}$$

Thus the estimates would be

$$\begin{aligned}\underline{b}&=(X_1'X_1)^{-1}X_1'\underline{y}_1\\ &=(X'X-X_2'X_2)^{-1}X_1'\underline{y}_1\\ &=\left[\left((I-X_2'X_2(X'X)^{-1}\right)X'X\right]^{-1}X_1'\underline{y}_1\\ &=(X'X)^{-1}(I-X_2'X_2(X'X)^{-1})^{-1}X_1'\underline{y}_1\end{aligned}$$

Given, $(I+AB)^{-1}=I-A(I+BA)^{-1}B$, by Tocher (1952).

Let $A=-X_2'$ and $B=X_2(X'X)^{-1}$, then

$$\begin{aligned}\underline{b}&=(X'X)^{-1}\left[I+X_2'(I-X_2(X'X)^{-1}X_2')^{-1}X_2(X'X)^{-1}\right]X_1'\underline{y}_1\\ &=(X'X)^{-1}\left[I+X_2'MX_2(X'X)^{-1}\right]X_1'\underline{y}_1\end{aligned}$$

where $M=(I-X_2(X'X)^{-1}X_2')^{-1}$

After some algeara, the expected values for the missing observations are seen to be

$$\hat{\underline{y}}_2 = X_2\underline{b} = MX_2(X'X)^{-1}X_1'\underline{y}_1$$

Since $$X_1'\underline{y}_1 = (X_1', X_2')\begin{pmatrix}\underline{y}_1\\ \underline{0}\end{pmatrix} = X'\underline{y}^{(0)}$$

it follows that $\hat{\underline{y}}_2 = MX_2\underline{b}_0$.

where $\underline{b}_0$ is the estimate of $\underline{\beta}$ obtained from the data assuming the missing observations have zero values.

Now we estimate the missing values $\underline{y}_2$ by choosing them in such a way that the residual sum of squares is minimized with respect to those values.

Residual sum of squares are

$$S^2 = \underline{y}'\underline{y} - \underline{b}'X'\underline{y} = \underline{y}'\underline{y} - \underline{y}'X(X'X)^{-1}X'\underline{y}$$
$$= \underline{y}'(I - X(X'X)^{-1}X')\underline{y} = \underline{y}'H\underline{y}$$

where $$H = I - X(X'X)^{-1}X'$$

$$\frac{\partial}{\partial \underline{y}}(S^2) = 2H\underline{y}$$

Let $$\frac{\partial}{\partial \underline{y}}(S^2) = \underline{0},\quad H\underline{y} = \underline{0}$$

Since $\frac{\partial}{\partial y_i}\underline{h}_i'\underline{y} = h_{ii} > 0$, because H is positive definite when $S^2 > 0$, this equation does give minimum.

Let $X' = [X_1', X_2']$, then

$$X'X = [X_1', X_2']\begin{pmatrix}X_1\\ X_2\end{pmatrix} = X_1'X_1 + X_2'X_2 \text{ and}$$

$$H = \left(\begin{array}{c|c} I - X_1(X'X)^{-1}X_1' & -X_1(X'X)^{-1}X_2' \\ \hline -X_2(X'X)^{-1}X_1' & I - X_2(X'X)^{-1}X_2' \end{array}\right)$$

Thus if $\underline{y}_1' = (y_1, \cdots, y_{N-h})$, is the vector of observed values and $\underline{y}_2' = (y_{N-h+1}, \cdots, y_N)$ is the vector of missing values, we obtained the estimates $\hat{\underline{y}}_2$ from

$$-X_2(X'X)^{-1}X_1'\underline{y}_1 + (I - X_2(X'X)^{-1}X_2')\hat{\underline{y}}_2 = \underline{0}$$

which implies that

$$\hat{\underline{y}}_2 = (I - X_2(X'X)^{-1}X_2')^{-1}X_2(X'X)^{-1}X_1'\underline{y} = MX_2(X'X)^{-1}X_1'\underline{y}$$

where $$M = (I - X_2(X'X)^{-1}X_2')^{-1}$$

3. Analysis of Variance

The correct analysis of variance table is

Source	S. S	D. F	M. S
Coefficients $\underline{b}$	$\underline{b}'X_1'\underline{y}_1$	k+1	
Residual	Difference	$N-h-k-1$	s^2
Total	$\underline{y}_1'\underline{y}_1$	$N-h$	

in which only the observations $\underline{y}_1$ appear and where $k+1$ is the number of coefficients estimated. The correct variance-covariance matrix of the $\underline{b}$ coefficients is

$$\hat{Var}(\underline{b}) = (X_1'X_1)^{-1} s^2 = (X'X)^{-1}\left[I + X_2'MX_2(X'X)^{-1}\right]s^2$$

4. A Single Missing point in 2^n Factorial Designs

In this section we shall discuss some of the consequences of one or two missing points in a 2^n factorial and also in some fractions. We shall see that the loss of a single point can cause the efficiency of the design to decrease by 50 %, and that the loss of certain pairs of points from a minimal fraction of resolution Ⅳ reduces it to a fraction of resolution Ⅱ.

It will be assumed throughout that the several factors have coordinates $x_i = \pm 1$; the tern "effect" will be used to denote either a main effect or an interaction, and by an estimate of an effect we shall mean an estimate of the corresponding regression coefficient, $\underline{\beta}$; for the complete factorial $V = (\hat{\underline{\beta}}) = \sigma^2/2^n$.

If P and Q are effects in the same alias set, then PQ is a defining contrast, and, if Q is suppressed, P is estimable from the half replicate defined either by $I = +PQ$, or $I = -PQ$, in which it is aliased with Q. If there are p suppressed effects in the same alias set as P we obtain p estimates of β_p from different half replicates.

We shall make frequent use of a theorem (John (1969, 1971)) which shows that the least squares estimate $\hat{\beta}_p$ has the following properties:

(1) $\hat{\beta}_p$ is the simple average of the p estimates of P from the half replicates,

(2) $V(\hat{\beta}_p) = 2^{-n}\sigma^2(p+1)/p$,

(3) if R is another effect in the same alias set as P, then Cov $(\hat{\beta}_p, \hat{\beta}_r) = 2^{-n}\sigma^2/p$: if R is in a different set from P, then Cov $(\hat{\beta}_p, \hat{\beta}_r) = 0$.

Suppose that one point in a 2^n factorial design is missing and that we desire to analyze the incomplete design. One approach is to compute a "missing plot" value x, for the missing point. A reasonable procedure for doing this is to suppress one effect, usually the n factor interaction, in the model. and to choose x so as to make the contrast for this effect zero. This procedure is equivalent to the method used in randomized complete block designs of choosing x so as to minimize the sum of squares for error, and it gives the least squares estimates of the remaining effects from the $2^n - 1$ actual points.

In practice the simplest way to do this is to carry out Yates' algorithm with a zero entry for the missing point and then add $\pm x$ as appropriate in the last column. We illustrate this in the first of the examples given in table.

Table- Examples of 2^4 designs. (Yates' algorithm)

Trt. comb	Data	Contrast	Example 1		Example 2		Contrast Identification
			Data	Contrast	Data	Contrast	
(1)	15	336	15	314+x	15	290+x+y	M
d	26	12	26	−10+x	26	16−x+y	D
c	18	−32	18	−10−x	18	−28−x+y	C
cd	21	32	21	54−x	y	−14+x+y	CD
b	28	4	28	−18+x	28	50−x −y	B
bd	22	− 8	x	−30+x	22	−12+x −y	BD
bc	11	−32	11	−10−x	11	−36+x−y	BC
bcd	19	24	19	46−x	19	70−x−y	BCD
a	25	16	25	38−x	x	12+x−y	A
ad	17	−20	17	2 −x	17	26−x−y	AD
ac	20	12	20	−10+x	20	58−x−y	AC
acd	24	20	24	− 2 +x	24	16+x−y	ACD
ab	29	4	29	26−x	29	8 −x+y	AB
abd	22	16	22	38−x	22	−30+x+y	ABD
abc	16	4	16	−18+x	16	−42+x +y	ABC
abcd	23	−20	23	−42+x	23	−16−x+y	ABCD

If we regard the missing point as an omitted 2^{n-n} fraction, applying the theorem with $p=1$ shows that for all effects, except the one suppressed, $V(\hat{\beta})=\sigma^2/2^{n-1}$; this is a loss of 50% in variance efficiency.

In Example 1 in the Table, bd is missing.

Let $\frac{\partial}{\partial x}(-30+x)^2=0$, then the estimated missing value is $x=30$.

Hence,

$$\hat{\beta}_1=(38-30)/8=1 \tag{1}$$

$\hat{\beta}_1$ from the single half replicate defined by $I=+ABD$ is

$$\hat{\beta}_1=(1/4)\ (abcd+abd+ac+a-bc-b-cd-d)=1. \tag{2}$$

This is the same as the (1)

$$V(\hat{\beta}_1)=\sigma^2/2^{n-1}=\sigma^2/8 \tag{3}$$

The experimenter who regards a loss of 50% in variance efficiency as unacceptable may be able to cut his loss of variance efficiency by deciding to suppress all the interactions involving $(n-1)$ factors as well as the n factor interaction. He will then have

$$p=\binom{n}{n-1}+1=n+1$$

and

$$V(\hat{\beta})=2^{-n}\sigma^2(n+2)/(n+1)$$

The new value of the variance efficiency will be $(n+1)/(n+2)$. The "missing plot" value x is again chosen so as to minimize the sum of squares for error, which is now the sum of the squares of each of the $n+1$ suppressed interactions. In our 2^4 example we should now suppress BCD, ACD, ABD, ABC and ABCD and minimize

$$S=[BCD]^2+[ACD]^2+[ABD]^2+[ABC]^2+[ABCD]^2$$
$$=(46-x)^2+(-2+x)^2+(38-x)^2+(-18+x)^2+(-42+x)^2$$

Let $\frac{\partial S}{\partial x}=0$, then, $x=29.2$

The least squares estimate of β_1 becomes $\hat{\beta}_1^*=(38-29.2)/8=1.1$ (4)

The estimate of β_1 from the half replicate defined by

$$I=+BC\ (A\sim ABC) \tag{5}$$

is

$$(abcd+abc+ad+a-bcd-bc-d-(1))/4=2.5 \tag{6}$$

The estimate of β_1 from the half replicate defined by

$$I=-BD\ (A\sim ABD) \tag{7}$$

is

$$(abc+ab+acd+ad-bc-b-cd-d)/4=0 \tag{8}$$

The estimate of β_1 from the half replicate defined by

$$I=+CD\ (A\sim ACD) \tag{9}$$

is

$$(abcd+ab+acd+a-bcd-b-cd-(1))/4=4.5 \tag{10}$$

The estimate of β_1 from the half replicate defined by

$$I=-ABCD\ (A\sim BCD) \tag{11}$$

is

$$(abc+abd+acd+a-bcd-b-c-d)/4=-1 \tag{12}$$

The estimate of β_1 from the half replicate defined by

$$I=+BCD\ \ (A\sim ABCD) \tag{13}$$

is

$$(abcd+ab+ac+ad-bcd-b-c-d)/4=-0.5 \tag{14}$$

From (6), (8), (10), (12), and (14), average $\hat{\beta}_1^*$ is

$$\hat{\beta}_1^* = (2.5 + 0 + 4.5 + (-1) + (-0.5)) / 5 = 1.1 \quad (15)$$

That is the same as the (4).

The variance of $\hat{\beta}_1^*$ is

$$V(\hat{\beta}_1^*) = 2^{-n}\sigma^2 (n+2) / (n+1) = 3\sigma^2/40 \quad (16)$$

From (3) and (16), the relative efficiency of $\hat{\beta}_1^*$ to $\hat{\beta}_1$ is 5/3.

5. Two Missing Points in 2^n Factorial Designs

Here we omit a $2^{n-(n-1)}$ fraction and there are two alias sets. In Example 2, we have assumed that a and cd are both lost; two missing values are calculated, x for a and y for cd. Yates' algorithm is performed with zero for a and cd and $\pm x$, $\pm y$ are added as appropriate. Again suppressing all five interactions, we choose x and y so as to minimize

$$S = (70-x-y)^2 + (16+x-y)^2 + (-30+x+y)^2 + (-42+x+y)^2 + (-16-x+y)^2$$

Let $\frac{\partial S}{\partial x} = 0$, and $\frac{\partial S}{\partial y} = 0$, then we find $x = 47/3$, and $y = 95/3$.

Hence,

$$\hat{\beta}_1 = (12+x-y) / 8 = -0.5$$
$$\hat{\beta}_2 = (50-x-y) / 8 = 1/3$$
$$\hat{\beta}_3 = (-28-x+y) / 8 = -1.5$$
$$\hat{\beta}_4 = (16-x+y) / 8 = 4$$

We introduce finding the Var$(\hat{\beta})$ that was studied by John (1979): Let us assume that we are still prepared to suppress all the interactions with n or $(n-1)$ factors. Suppose that there are t factors that change levels in the two missing points and $s = (n-t)$ that do not. The defining contrast subgroup contains (i) the subgroup of 2^s elements generated by the s unchanged factors and (ii) the set of 2^{t-1} elements consisting of all interactions with an even number of letters from the t changing factors.

If t is even, (ii) contains the product of all t changing factors and, hence, the n factor interactions, but no product of $t-1$ of the changing factors; it also contains s of the $(n-1)$ factor interactions. Estimated effects in the subgroup have $V(\hat{\beta}) = 2^{-n}\sigma^2 (s+2) / (s+1)$. The other alias set contains the remaining t suppressed interactions and $V(\hat{\beta}) = 2^{-n}\sigma^2 (t+1) / t$. If t is odd, the situation is reversed.

In this Example the alias sets are I, B, AC, AD, CD, ABC, BCD, ABD,

and A, C, D, AB, BC, BD, ACD, ABCD ; again A is estimable from the half replicate in which A is aliased with ABCD ($I=+BCD$). Hence $V(\hat{\beta}_1)=V(\hat{\beta}_3)=V(\hat{\beta}_4)=3\sigma^2/32$ and $V(\hat{\beta}_2)=\sigma^2/12$.

References

1. Allen, F. E and Wishart, J. (1930). *A Method of Estimating the Yield of a Missing Plot in Field Experimental Work.* J. Agric. Sci., 20, 399-406.
2. Draper, N. R. and Stoneman, D. M. (1963). *Estimating Missing Values in Unreplicated Two-level Factorial and Fractional Factorial Designs.* University of Wisconsin Technical Report No. 20.
3. John, P. W. M. (1969). *Some-Non-orthogonal Fractions of 2^n Designs.* J. Roy. Statist. Soc., B, 31, 270-275.
4. John, P. W. M. (1971). *Statistical Design and Analysis of Experiments.* New York : The Macmillan Company.
5. John. P. W. M. (1979) *Missing points in 2^n and 2^{n-k} Factorial Designs.* Technometrics, Vol. 21, No. 2. 225-228.
6. Tocher, K. D., (1952). *The design and Analysis of Block Experiments.* Journal of the Royal Statistical Society, Series B, Vol. 14, 45-100.
7. Wilkinson, G, N. (1958). *Estimation of Missing Values for the Analysis of Incomplete Data.* Biometrics, 14, 257-286.
8. Wilkinson, G. N. (1970). *A General Recursive Procedure for Analysis of Variance.* Biometrika, 57, 19-46.
9. Yates, F. (1933). *The Analysis of Replicated Experiments when the Field Results are Incomplete.* Emp. J. Exp. Agric., 1, 129-142.